



Development and validation of a simulation workload measure: the simulation task load index (SIM-TLX)

David Harris¹ · Mark Wilson¹ · Samuel Vine¹

Received: 13 April 2019 / Accepted: 11 December 2019
© The Author(s) 2019

Abstract

Virtual reality (VR) simulation offers significant potential for human factors training as it provides a novel approach which enables training in environments that are otherwise dangerous, impractical or expensive to simulate. While VR training has been adopted in many environments, such as heavy industry, surgery and aviation, there remains an inadequate understanding of how virtual simulations impact cognitive factors. One such factor, which needs careful consideration during the design of VR simulations, is the degree of mental or cognitive load experienced during training. This study aimed to validate a newly developed measure of workload, based on existing instruments (e.g. the NASA-TLX), but tailored to the specific demands placed on users of simulated environments. While participants completed a VR puzzle game, a series of experimental manipulations of workload were used to assess the sensitivity of the new instrument. The manipulations affected the questionnaire subscales (mental demands; physical demands; temporal demands; frustration; task complexity; situational stress; distraction; perceptual strain; task control; presence) as predicted in all cases ($ps < .05$), except for presence, which displayed little relationship with other aspects of task load. The scale was also found to have good convergent validity with an alternate index of task load. The findings support the sensitivity of the new instrument for assessing task load in virtual reality. Overall, this study contributes to the understanding of mental workload in simulated environments and provides a practical tool for use in both future research and applications in the field.

Keywords Workload · Cognitive load · Learning · Virtual reality · Training

1 Introduction

Simulation is well established as an effective method for human factors training, particularly in environments such as surgery and aviation (Stevens and Kincaid 2015; Sutherland et al. 2006). Recently there has been increasing interest in the use of virtual reality (VR) technologies for simulation training, as VR affords innovative training methods that can be applied to otherwise dangerous or impractical environments. VR allows the user to interact with a simulation of some real environment, in real time, using their own senses and motor skills (Burdea and Coiffet 2003). A number of

studies have shown beneficial effects of VR training for sport and exercise (Gray 2017; Neumann et al. 2018), surgery (Hashimoto et al. 2018), equipment operators (Dunston et al. 2014), rehabilitation (Adamovich et al. 2009) and mental health (Krijn et al. 2004). Currently, however, there is limited understanding of how VR environments impact the cognitive processes of the user. For instance, it is unclear whether the cognitive demands imposed by a virtual task are equivalent to that of the real world. As cognitive load plays an important role in informational learning and skill acquisition (Kirschner 2002; Renkl and Atkinson 2003; van Gog et al. 2005), this lack of understanding could limit the training effectiveness of VR. Consequently, validated methods for assessing workload in simulators, and virtual reality simulators in particular, are required. This study aims to provide a validated method for assessing workload to aid future design and assessment of simulated environments.

Workload is a multifaceted construct, which results from the interaction of the intrinsic demands of the task and the conditions under which the task is performed, as well as the

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10055-019-00422-9>) contains supplementary material, which is available to authorized users.

✉ David Harris
d.j.harris@exeter.ac.uk

¹ University of Exeter, Exeter, UK

behaviour, skill and perceptions of the performer (DiDomenico and Nussbaum 2008). Cognitive load theory (CLT; Sweller 1999) describes mental workload as a result of the number of informational units that must be held in working memory—the domain general limited capacity information store (Pollock et al. 2002). CLT suggests that an optimal level of load exists for the purpose of informational learning, due to effects on engagement and attention. It is, therefore, the goal of the instructor to optimise load for the learner, by (1) providing a primary task load that is sufficient but not excessive and (2) by removing sources of load that are extraneous to the target task. The literature describing the challenge point framework (see Guadagnoli and Lee 2004) suggests that an optimal load is also best for motor learning and should be balanced based on the demands of the task and the (developing) skills of the learner.

The existing methods for assessing workload or effort can be broadly categorised into subjective and objective methods. Psychophysiological measures such as heart rate and heart rate variability (consistency in beat-to-beat interval) provide an online, objective measure of effort, but can also be influenced by physiological demands (Mulder et al. 2004). Dual tasking also provides an objective index of mental effort (Steed et al. 2016), which, based on the premise of limited working memory capacity, aims to assess the extent of spare capacity left by the primary task, through performance on a concurrent secondary task. While some researchers suggest this method to be a more reliable and direct measure of cognitive load (Brunken et al. 2003), it ignores the important subjective experience of workload, which can only be assessed using self-report. Perhaps the most widely used self-report measure of workload in human factors research is the NASA-task load index (NASA-TLX; Hart and Staveland 1988). The NASA-TLX assesses workload along six dimensions: mental demand; physical demand; temporal demand; own performance; effort; and frustration. The NASA-TLX was designed to provide better diagnosticity (i.e. to discriminate between different *types* of workload) by assessing multiple sources of load, rather than a unidimensional construct, but as it was designed for pilots (space flight), it may not reflect the unique demands posed by some tasks. Accordingly, it has been adapted for specific environments, such as the surgery task load index (SURG-TLX; Wilson et al. 2011) and the driving activity load index (DALI; Pauzie 2008). As VR simulations present a distinct, novel environment, which poses unique usability challenges, the existing measures may not address all relevant sources of workload.

Additional factors which must be considered when extending the current methods of assessing workload to VR include: (1) degree of immersion, (2) perceptual difficulties and (3) novel methods of controlling the environment. The level of *immersion* and *presence* provided by VR is proposed

as a particular benefit of virtual training, although immersion may be moderated by individual difference variables (Shin 2018). Immersion refers to the sensations induced by the technologies creating the VE, while presence refers to the user experience that they have left the real world and are now present in the virtual world (Mestre et al. 2006). The high degree of immersion provided by VR seems to be important for increased engagement in virtual training (Stanney et al. 2003), but it is unclear how it impacts workload. Potentially the immersive nature of VR may reduce distractions and the need for effortful focus (Wickens 1992). A second factor germane to VR is the perceptual difficulties resulting from virtual presentation of visual stimuli, such as mismatched motion and inadequate rendering rates which contribute to simulator sickness (Somrak et al. 2019). Finally, navigation in VR places new demands on the user, meaning trainees may have to learn to control the VR environment before they can even begin to access the benefits of the simulation for training the target task. These challenges are unique to VR and may significantly impact workload and learning and hence should be considerations for new methods of quantifying load in VR.

Although previous studies have attempted to gauge cognitive load during VR tasks (Bharathan et al. 2013; Frederiksen et al. 2019; Naismith et al. 2015), they fall short by using overly simple unidimensional measures, instruments that do not address the unique sources of load posed by virtual simulation (e.g. the NASA-TLX), or by ignoring multiple sources of load. Therefore, the purpose of the present study was to address these shortcomings by designing and validating a task load index, aimed primarily at virtual reality simulations but relevant to a range of simulated environments. Adopting the approach of Wilson et al. (2011) in developing the SURG-TLX, this study aimed to (1) develop a task load index specific to virtual reality simulation and (2) demonstrate its validity through experimentally manipulating relevant sources of workload.

2 Methods

2.1 Scale development

As well-validated task load instruments already exist in the form of the NASA-TLX (Hart and Staveland 1988) and SURG-TLX (Wilson et al. 2011), the new instrument was designed to maintain a similar overall structure, but to be more applicable to the demands of simulated environments. The scale was developed by selecting the relevant items from the existing instruments and adding the simulation-specific scales. The new scale adopted the dimensions *mental*, *physical*, *temporal* and *frustration* demands from the NASA-TLX, as they were still highly applicable to VR

simulations. Additionally, the dimensions *complexity*, *stress* and *distractions* were added from the surgery-specific version (SURG-TLX), as these captured additional sources of load not present in the NASA-TLX. Next, new items were added which were specific to workload in simulations. As discussed previously, VR poses a challenge in terms of perceptual demands and demands from controlling the environment. Hence, *perceptual strain* and *task control* scales were added to address these issues, using the same format at the questions from the original NASA version. Finally, a scale relating to *presence* was added, as feeling present in the virtual world is an important part of usability and training effectiveness of VR. In general, immersion refers to the capabilities of the technology in producing the simulated environment, while presence refers to the feeling of ‘being there’ created by the technology (Mestre et al. 2006). However, as these terms are often used interchangeably outside of academic discourse (and participants were unlikely to be aware of the subtle difference), both terms were provided in the questionnaire to aid the understanding of the participant. Therefore, the proposed dimensions for the SIM-TLX were:

1. Mental demands—How mentally fatiguing was the task?
2. Physical demands—How physically fatiguing was the task?
3. Temporal demands—How hurried or rushed did you feel during the task?
4. Frustration—How insecure, discouraged, irritated, stressed or annoyed were you?
5. Task complexity—How complex was the task?
6. Situational stress—How stressed did you feel while performing the task?
7. Distraction—How distracting was the task environment?
8. Perceptual strain—How uncomfortable/irritating were the visual and auditory aspects of the task?
9. Task control—How difficult was the task to control/navigate?
10. Presence—How immersed/present did you feel in the task?¹

2.2 Participants

Thirty-seven student participants (20 male, mean age = 20.5 years, SD = 3.5) volunteered to take part in the study. Participants were recruited from an undergraduate population using convenience sampling, through poster advertisements and word of mouth. This research complied

¹ While the term ‘immersion’ was added to the question to aid understanding, this item is aimed at the user experience of being in the environment.



Fig. 1 Virtual reality block stacking game

with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Exeter. All subjects provided written informed consent before testing.

2.3 Materials and task

2.3.1 Task

The experimental task was a block stacking puzzle game, often known as ‘Jenga’, which requires the player to remove blocks from the body of a tower, one at a time, and replace them on the top until the tower falls. The block stacking game was chosen as it is a perceptual-motor task, which also provides some cognitive challenge (i.e. game strategy). Additionally, the task could be easily completed in both real and virtual worlds in a laboratory setting. The virtual version of the task used an application developed in Unity (2018.2.16f1) and was displayed using an HTC Vive head mounted display (Fig. 1). Graphics displayed in the Vive were generated on a 3XS laptop running Windows 10 with an Intel i7 processor and Nvidia GeForce GTX 1070 graphics card. Players interacted with blocks using the HTC Vive hand-held controller. Blocks could be grasped by squeezing buttons on the side of the handle, which was controlled using the dominant hand.

2.3.2 SIM-TLX

In line with the original NASA-TLX and the SURG-TLX, a two-part evaluation was used to complete the SIM-TLX. Following each completion of the task, participants rated the level of demand on ten scales reflecting each of the

workload dimensions on a 21-point Likert scale, anchored between low and high (Part 1; see supplementary materials for the SIM-TLX). Additionally, participants indicated the relative importance of each workload dimension for the task, by making a series of 45 pairwise comparisons between dimensions (Part 2). A weight is assigned based on the cumulative score achieved for each dimension after the comparisons, which is multiplied with the Likert scale score to compute the workload score for each dimension (scores range from 0 to 189). For example, a weight score of 2 and a rating of 10 would equate to a workload score of 20. A total (unidimensional) workload score can also be calculated by aggregating the scores from the ten dimensions.

2.3.3 Additional measures

In order to further validate the SIM-TLX, we aimed to assess convergent and divergent validity, in order to establish construct validity of the SIM-TLX (Campbell and Fiske 1959). Convergent validity relates to the strength of the relationship between two methods intending to measure the same underlying construct. Divergent validity indicates that the construct you are measuring is conceptually distinct from another, related, construct.

To assess convergence with a rating scale of mental effort, we used the rating scale of mental effort (RSME, Zijlstra 1993). The RSME is an expedient unidimensional visual analogue scale, ranging from 0 to 150, with descriptors along the scale, such as ‘absolutely no effort’, ‘considerable effort’ and ‘extreme effort’. The RSME has been widely used and has been found to be reliable over repeated administrations in laboratory ($r=0.88$) and work settings ($r=0.78$; Zijlstra 1993). If a strong correlation were to be observed between the SIM-TLX and the RSME, it would indicate that a similar underlying construct was being measured (i.e. convergent validity).

To assess divergent validity, a measure was chosen that reflected a construct related to, but conceptually distinct from, workload as assessed by the SIM-TLX (Campbell and Fiske 1959). The enjoyment subscale of the Intrinsic Motivation Inventory (IMI; McAuley et al. 1989) was chosen, as workload was expected to relate to enjoyment (in a negative or inverted U pattern), but should remain conceptually distinct from it, and as such should not show a high correlation with enjoyment.

2.4 Procedure

Participants attended the laboratory on one occasion for approximately 60 min. Participants first completed the block stacking task in the real world, where they played the game (with no opponent) for 2 min. They then completed parts 1

and 2 of the SIM-TLX, IMI enjoyment scale and the RSME. Participants were then allowed 2 min to familiarise themselves with the VR environment and the controls before they completed the VR task under the five experimental conditions. The order of conditions was counterbalanced using a Latin squares design. After each condition, participants completed the questionnaires (SIM-TLX part 1, IMI enjoyment scale and RSME). After completing all 5 conditions, participants completed part 2 of the SIM-TLX in relation to all the VR tasks.

Conditions:

1. Control—participants were instructed to play the game for 2 min.
2. Multitasking—participants additionally counted backwards in sevens from 737, as in Wilson et al. (2011).
3. Time restriction—participants were asked to move 3 bricks in just 1 min (known to be a challenging target, based on pilot testing) and were informed when 30 s had passed and when only 5 s remained.
4. Stress—participants were read a script designed to induce social evaluative pressure, as in Wilson et al. (2011). The script informed participants that their score within the 2 min time limit would be entered onto a leader board for all participants to see and that the winner would receive a prize. They were also told that their performance in the pre-trial practice was poor and that they needed to improve.
5. Disruption—participants played the game for 2 min, but with a degraded visual display and impaired use of the controllers. A plastic insert was placed into the headset to reduce visual acuity and degrade the display, and the controllers were fixed to the users hand to restrict movement, so had to be used in an inverted orientation, which made control more difficult.

2.5 Data analysis

A workload score was computed for each dimension, by weighting the scores on the Likert scales of each workload dimension with the perceived importance based on the paired comparisons. A total workload score was calculated by summing the component dimensions. Missing pairwise weighting data for 1 participant was imputed based on predictive mean matching, using the R package ‘mice’ (van Buuren and Groothuis-Oudshoorn 2011). Statistical analyses were performed using R (v1.0.143, R Core Team 2017). The effect of condition on scale scores was analysed with a linear mixed effects model, using the lme4 package (Bates et al. 2014). Significant effects were explored using Bonferroni–Holm-corrected contrasts. Raw data and analysis scripts (as well as the full SIM-TLX scale) are available from the Open Science Framework (<https://osf.io/p6de4/>).

2.6 Hypotheses

It is predicted that relative to other conditions:

- H₁:** Mental demands will be increased in the multitasking condition.
- H₂:** Physical demands will be increased in the disruption condition.
- H₃:** Temporal demands will be increased in the time restriction condition.
- H₄:** Frustration demands will be increased in time restriction, disruption and multitasking conditions.
- H₅:** Task complexity demands will be increased in the multitasking condition.
- H₆:** Situational stress demands will be increased in the stress condition.
- H₇:** Distraction demands will be increased in the multitasking and disruption conditions.
- H₈:** Perceptual strain demands will be increased in the disruption condition.
- H₉:** Task control demands will be increased in the disruption condition.
- H₁₀:** Presence will be impaired by all conditions apart from control.

3 Results

3.1 Mental demands

A linear mixed effects model, with random intercepts for participants, was run to examine the effect of condition on the mental demands scale (and each subsequent subscale). The overall model predicting mental demands score had a total explanatory power (conditional R^2) of 70.80%, in which the fixed effect of condition explained 16.99% of the variance. The model's intercept was at 22.51 (SE=5.70, 95% CI [11.34, 33.68]). The effect of condition was statistically significant, $F(5,180) = 25.72$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, mental demands were significantly increased in the multitasking condition relative to other conditions ($ps < .001$) (Fig. 2). A complete list of all t tests is available with the supplementary materials (<https://osf.io/p6de4/>).

3.2 Physical demands

The overall model predicting physical demands score had a total explanatory power of 77.82%, in which the fixed effect of condition explained 5.30% of the variance. The model's intercept was at 4.51 (SE=3.29, 95% CI [-1.96, 10.99]). The effect of condition was statistically significant,

$F(5,180) = 10.56$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, physical demands were higher in the disruption condition than in all other conditions ($ps < .05$), apart from stress ($p = .08$) (Fig. 2).

3.3 Temporal demands

The overall model predicting temporal demands score had a total explanatory power of 65.55%, in which the fixed effect of condition explained 25.69% of the variance. The model's intercept was at 10.11 (SE=4.33, 95% CI [1.65, 18.57]). The effect of condition was statistically significant, $F(5,180) = 32.96$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, temporal demands were increased in the time restriction condition relative to all other conditions ($ps < .001$), apart from stress ($p = .08$) (Fig. 2).

3.4 Frustration

The overall model predicting frustration score had a total explanatory power of 69.82%, in which the fixed effect of condition explained 10.18% of the variance. The model's intercept was at 15.54 (SE=6.84, 95% CI [2.14, 28.94]). The effect of condition was statistically significant, $F(5,180) = 14.90$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, frustration was increased in the time restriction and multitasking conditions, relative to real-world and control conditions ($ps < .02$). Frustration was also increased in the disruption condition, relative to the real world ($p < .001$), but not control ($p = .54$). Additionally, frustration scores were also increased in the stress condition, relative to both real-world and control conditions ($ps < .02$) (Fig. 2).

3.5 Task complexity

The overall model predicting task complexity score had a total explanatory power of 78.45%, in which the fixed effect of condition explained 19.47% of the variance. The model's intercept was at 23.81 (SE=6.47, 95% CI [11.10, 36.52]). The effect of condition was statistically significant, $F(5,180) = 39.95$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, task complexity scores were increased in the multitasking condition relative to all other conditions ($ps < .009$) (Fig. 2).

3.6 Stress

The overall model predicting stress score had a total explanatory power of 63.67%, in which the fixed effect of condition explained 20.19% of the variance. The model's intercept was at 14.51 (SE=5.46, 95% CI [3.85, 25.18]).

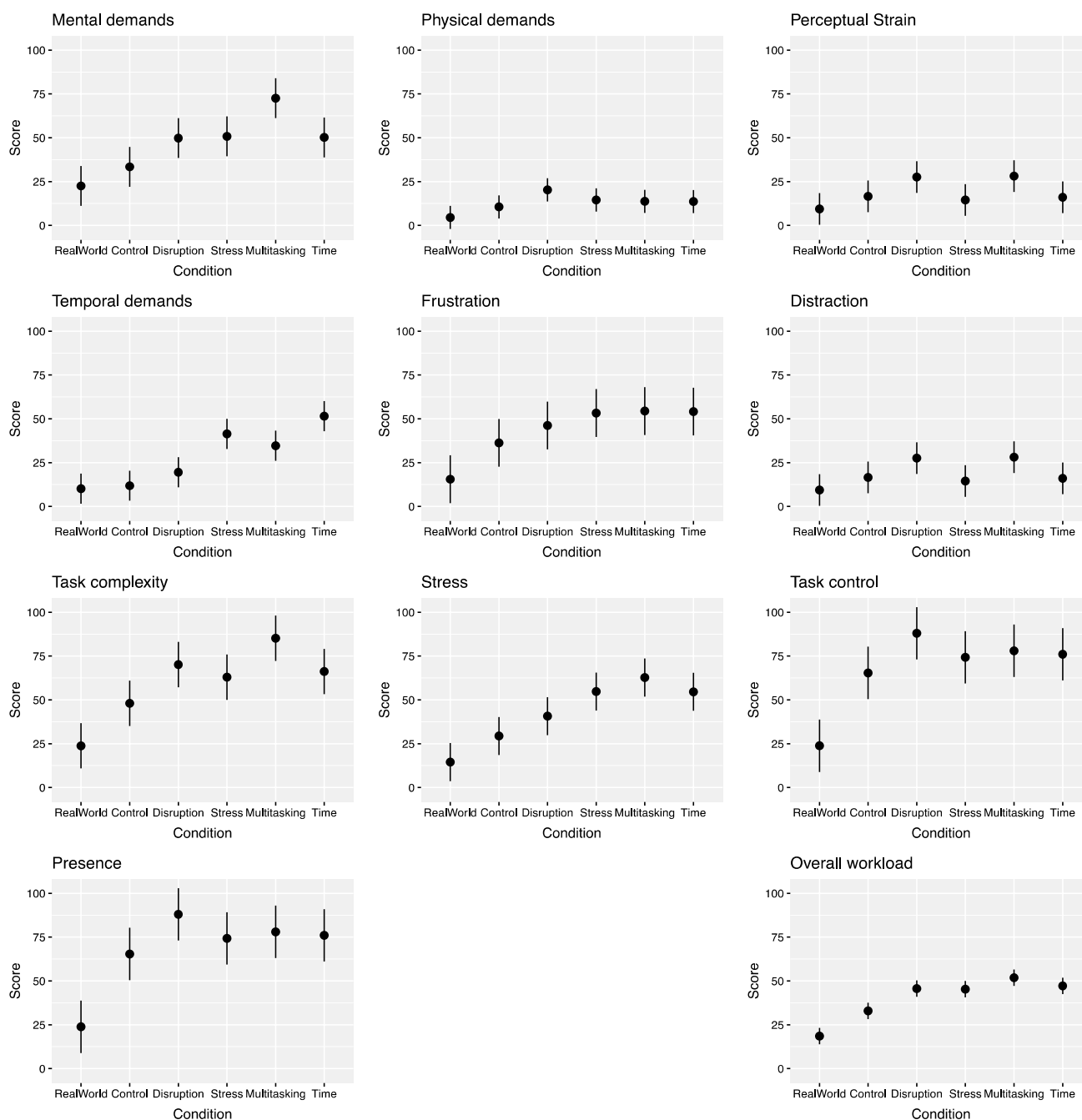


Fig. 2 Mean workload (and 95% CIs) for experimental conditions for all scales of the SIM-TLX

The effect of condition was statistically significant, $F(5,180) = 24.57$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, stress scores were increased in the social evaluative stress condition, relative to disruption, control and real-world conditions ($ps < .05$). Additionally, stress scores were also increased in multitasking and time pressure conditions, relative to real world, control and disruption ($ps < .05$). There was no difference in stress scores between multitasking, stress and time pressure conditions ($ps > .36$) (Fig. 2).

3.7 Distraction

The overall model predicting distraction score had a total explanatory power of 65.71%, in which the fixed effect of condition explained 5.92% of the variance. The model's intercept was at 9.35 (SE = 4.53, 95% CI [0.47, 18.23]). The effect of condition was statistically significant, $F(5,180) = 7.63$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, distraction scores

were increased in the multitasking and disruption conditions, relative to all other conditions ($ps < .05$) (Fig. 2).

3.8 Perceptual strain

The overall model predicting perceptual strain score had a total explanatory power of 62.25%, in which the fixed effect of condition explained 15.53% of the variance. The model's intercept was at 3.84 ($SE = 2.65$, 95% CI $[-1.34, 9.02]$). The effect of condition was statistically significant, $F(5, 180) = 18.18$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that, as predicted, perceptual strain scores were elevated in the disruption condition relative to all other conditions ($ps < .001$) except multitasking ($p = .10$) (Fig. 2).

3.9 Task control

The overall model predicting task control score had a total explanatory power of 80.45%, in which the fixed effect of condition explained 17.26% of the variance. The model's intercept was at 23.84 ($SE = 7.45$, 95% CI $[9.17, 38.50]$). The effect of condition was statistically significant, $F(5, 180) = 39.03$, $p < .001$. Bonferroni–Holm-corrected comparisons indicated that task control scores were elevated in the disruption condition compared to control and real-world conditions ($ps < .001$), but not compared to multitasking, stress or time ($ps > .07$) (Fig. 2).

3.10 Presence/immersion

The overall model predicting presence score had a total explanatory power of 81.51%, in which the fixed effect of condition explained 1.25% of the variance. The model's intercept was at 57.24 ($SE = 7.93$, 95% CI $[41.62, 72.87]$). The effect of condition was statistically significant, $F(5, 180) = 3.00$, $p = .01$. Bonferroni–Holm-corrected comparisons indicated that there were only small differences in presence scores across conditions (see Fig. 2). Presence only differed between time and real-world ($p = .006$) and real-world and stress ($p = .05$) conditions. There were no other differences between conditions ($ps > .22$).

3.11 Convergent validity and divergent validity

A linear mixed model, with random intercepts for participants, was used to assess convergent validity. The model, predicting SIM-TLX overall workload using RSME scores, had a total explanatory power of 79.61%. The model's intercept was at 8.91 ($SE = 2.17$, 95% CI $[4.63, 13.16]$). Within

this model, the effect of RSME was significant ($\beta = 0.50$, $SE = 0.025$, 95% CI $[0.45, 0.55]$, $t(214) = 20.27$, $p < .001$) and accounted for 60.05% shared variance between the workload measures. This indicates a high degree of convergence.

A linear mixed model, with random intercepts for participants, was also used to assess divergent validity. The model, predicting SIM-TLX overall workload using enjoyment scores from the IMI, had a total explanatory power of 29.70%. The model's intercept was at 35.32 ($SE = 6.46$, 95% CI $[22.62, 48.17]$). Within this model, the effect of enjoyment was not significant ($\beta = 0.46$, $SE = 0.58$, 95% CI $[-0.70, 1.61]$, $t(220) = 0.80$, $p = .21$) and there is only 0.29% shared variance, indicating the two constructs to be distinct.

3.12 Correlations

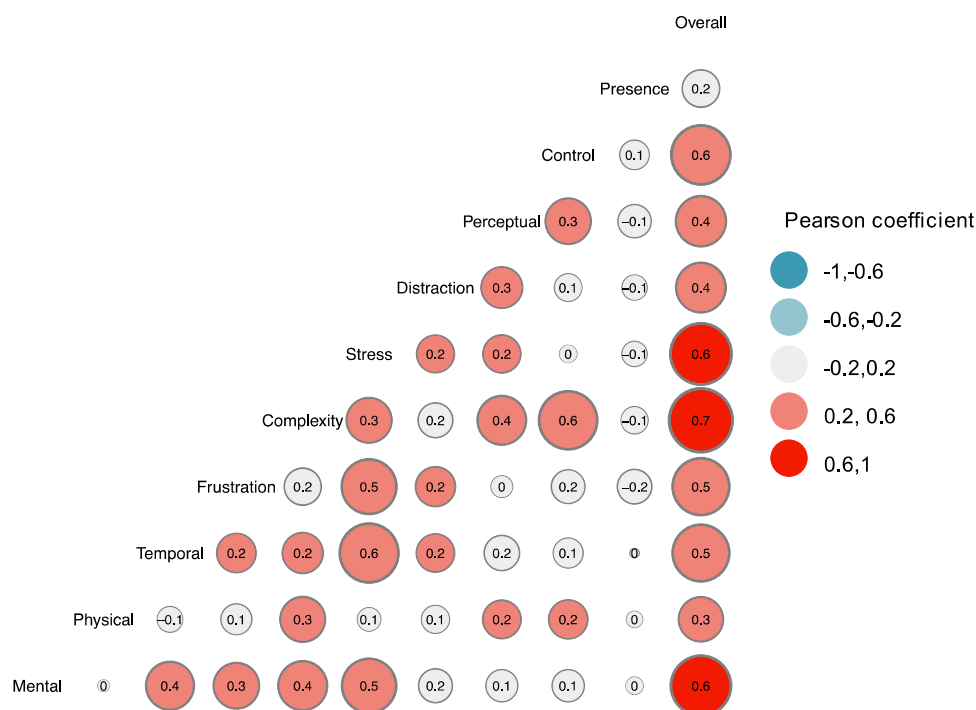
An exploratory analysis of the relationships between the scales of the SIM-TLX was run to understand how the different aspects of workload related to each other, and in particular the VR simulation-specific scales. The correlation matrix (Fig. 3) indicates that most scales showed moderate correlations with the overall measure, and with each other, as expected. The presence scale, however, shows the weakest relationship with overall workload and little to no relationship with other scales.

4 Discussion

The aim of this study was to explore the potential of a new multidimensional, simulation-specific, task workload measure (SIM-TLX). Virtual reality is becoming increasingly important for human factors training, but there is still limited understanding of how VR impacts cognitive processes, and therefore learning. Appropriate workload measures can help to improve VR and simulation training through optimising cognitive load for learning (Kirschner 2002). Workload measures are also of benefit during the design process, to ensure that simulated tasks challenge trainees in a way that is akin to 'real-world' tasks. Consequently, this paper makes an important step by designing and developing a novel tool that may be useful for assessing task load in simulated environments.

To validate the new SIM-TLX, we tested the sensitivity of the subscales to a range of manipulations, as in Wilson et al. (2011). The experimental manipulations were successful in effecting changes in the predicted scales in almost all cases, supporting the validity of the instrument. Mental demands were elevated when the VR task was supplemented with an additional mathematics task, and physical demands were increased when use of the VR controllers was impaired, although physical demands remained low overall. Temporal demands were notably increased in the time restriction

Fig. 3 Correlation matrix (with Pearson correlation coefficients) between SIM-TLX subscales and overall workload measure



condition, although they were also elevated in the social evaluative pressure condition. This is unsurprising, as in this condition participants were also made aware of the time limit. Frustration was increased in a number of conditions, and task complexity was clearly elevated in the multitasking condition, as expected. Stress scores were increased in the social evaluative stress condition, as well as time and multitasking conditions. Distraction and perceptual strain were affected by multitasking and disruption conditions as expected, and task control was impaired by the disruption condition, but only in comparison with control. Notably, there was little effect of any of the manipulations on reported presence/immersion in the task.

Presence was included in the simulation-specific workload measure, as presence or immersion is thought to be important for motivation and engagement with VR simulation training (Stanney et al. 2003). It has been suggested that presence in VR may reduce cognitive load, as less attentional effort has to be expended to consciously focus on the task (Wickens 1992), and Lackey et al. (2016) found that immersion correlated with additional mental demands. Alternatively, Rose et al. (2000) have suggested that VR might impose *additional* load when the user has to cope with mismatches of visual feedback and vestibular and proprioceptive feedback. The present results, however, suggest that presence may have no clear relationship with workload. There was little effect of the workload manipulations on presence, and the correlation matrix (Fig. 3) indicates that presence had little to no relationship with other aspects of workload. Certainly, the relationship between presence

and workload warrants further study, but based on these results presence should perhaps be treated as independent of workload. Consequently, we have excluded the presence scale from the final SIM-TLX and recommend that, when assessing the subjective experience of VR, it be considered a stand-alone construct (Usuh et al. 2000).

Drawing conclusions about the difference in workload between real and simulated tasks was not the primary aim of this study, but results indicated workload to be considerably higher in the virtual compared to the real block stacking task (similar results have been observed during surgical simulation; Frederiksen et al. 2019). As discussed previously, cognitive load plays an important role in optimising learning (Guadagnoli and Lee 2004; Kirschner 2002), but it is likely that the relationship between real and simulated tasks is very much specific to the task in question and dependent on the fidelity of the VR environment.

Overall, we found good support for the validity of the SIM-TLX, once the presence scale had been removed. While the results here support the sensitivity of the measure to the manipulations, additional validation is still required in other populations and tasks. The individual dimensions, other than presence, were affected by the manipulations in the manner expected and all showed moderate relationships with the rest of the workload scale. Workload scores suggested that physical demands were low in this task, but this scale has been retained as physical effort may be an important feature of other simulated tasks. Additionally, the overall scale showed convergent validity with a unidimensional measure of mental effort and divergent validity from an enjoyment

scale. In summary, although continued testing is required, these initial results support the validity of the final SIM-TLX (available from the Open Science Framework <https://osf.io/p6de4/>) as a multidimensional measure that can be employed to support the development and validation of simulated environments, and virtual reality training in particular.

5 Key points

- Workload plays a significant role in learning, but there is currently limited understanding of how simulated training affects the workload experienced by users.
- A simulation-specific, validated method of assessing load is needed to more effectively design and validate simulated environments.
- This study presents and validates a new measure, the SIM-TLX, for assessing workload in simulation and virtual reality.
- This new measure can be applied to research and the development and validation of simulators for human factors training.

Acknowledgements This work was supported by a Royal Academy of Engineering Postdoctoral Fellowship and an NVIDIA equipment grant awarded to D Harris.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adamovich SV, Fluet GG, Tunik E, Merians AS (2009) Sensorimotor training in virtual reality: a review. *NeuroRehabilitation* 25(1):29–44. <https://doi.org/10.3233/NRE-2009-0497>
- Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. Retrieved from <http://arxiv.org/abs/1406.5823>
- Bharathan R, Vali S, Setchell T, Miskry T, Darzi A, Aggarwal R (2013) Psychomotor skills and cognitive load training on a virtual reality laparoscopic simulator for tubal surgery is effective. *Eur J Obstet Gynecol Reprod Biol* 169(2):347–352. <https://doi.org/10.1016/j.ejogrb.2013.03.017>
- Brunken R, Plass JL, Leutner D (2003) Direct measurement of cognitive load in multimedia learning. *Educ Psychol* 38(1):53–61. https://doi.org/10.1207/S15326985EP3801_7
- Burdea GC, Coiffet P (2003) Virtual reality technology. Wiley, Hoboken
- Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 56(2):81–105
- DiDomenico A, Nussbaum MA (2008) Interactive effects of physical and mental workload on subjective workload assessment. *Int J Ind Ergon* 38(11):977–983. <https://doi.org/10.1016/j.ergon.2008.01.012>
- Dunston PS, Proctor RW, Wang X (2014) Challenges in evaluating skill transfer from construction equipment simulators. *Theor Issues Ergon Sci* 15(4):354–375. <https://doi.org/10.1080/1463922X.2011.624647>
- Frederiksen JG, Sørensen SMD, Konge L, Svendsen MBS, Nobel-Jørgensen M, Bjerrum F, Andersen SAW (2019) Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: a randomized trial. *Surg Endosc*. <https://doi.org/10.1007/s00464-019-06887-8>
- Gray R (2017) Transfer of training from virtual to real baseball batting. *Front Psychol*. <https://doi.org/10.3389/fpsyg.2017.02183>
- Guadagnoli MA, Lee TD (2004) Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *J Mot Behav* 36(2):212–224. <https://doi.org/10.3200/JMBR.36.2.212-224>
- Hart SG, Staveland LE (1988) Development of NASA-TLX (task load index): results of empirical and theoretical research. In: *Advances in psychology*, vol 52, pp 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hashimoto DA, Petrusa E, Phitayakorn R, Valle C, Casey B, Gee D (2018) A proficiency-based virtual reality endoscopy curriculum improves performance on the fundamentals of endoscopic surgery examination. *Surg Endosc* 32(3):1397–1404. <https://doi.org/10.1007/s00464-017-5821-5>
- Kirschner PA (2002) Cognitive load theory: implications of cognitive load theory on the design of learning. *Learn Instr* 12(1):1–10. [https://doi.org/10.1016/S0959-4752\(01\)00014-7](https://doi.org/10.1016/S0959-4752(01)00014-7)
- Krijn M, Emmelkamp PMG, Olafsson RP, Biemond R (2004) Virtual reality exposure therapy of anxiety disorders: a review. *Clin Psychol Rev* 24(3):259–281. <https://doi.org/10.1016/j.cpr.2004.04.001>
- Lackey SJ, Salcedo JN, Szalma JL, Hancock PA (2016) The stress and workload of virtual reality training: the effects of presence, immersion and flow. *Ergonomics* 59(8):1060–1072. <https://doi.org/10.1080/00140139.2015.1122234>
- McAuley E, Duncan T, Tammen VV (1989) Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis. *Res Q Exerc Sport* 60(1):48–58. <https://doi.org/10.1080/02701367.1989.10607413>
- Mestre D, Fuchs P, Berthoz A, Vercher JL (2006) Immersion et présence. In: Fuchs P, Moreau G, Berthoz A, Vercher JL (eds) *Le Traite de la realite Virtuelle*. Ecole Des Mines de Paris, Paris, pp 309–338
- Mulder JM, de Waard D, Brookhuis KA (2004) Estimating mental effort using heart rate and heart rate variability. In: Stanton N, Hedge A, Hendrick HW, Brookhuis KA, Salas E (eds) *Handbook of ergonomics and human factors methods*. Taylor and Francis, London. <https://doi.org/10.1201/9780203489925-30>
- Naismith LM, Cheung JJH, Ringsted C, Cavalcanti RB (2015) Limitations of subjective cognitive load measures in simulation-based procedural training. *Med Educ* 49(8):805–814. <https://doi.org/10.1111/medu.12732>
- Neumann DL, Moffitt RL, Thomas PR, Loveday K, Watling DP, Lombard CL, Antonova S, Tremeer MA (2018) A systematic review of the application of interactive virtual reality to sport.

- Virtual Real 22(3):183–198. <https://doi.org/10.1007/s10055-017-0320-5>
- Pauzie A (2008) A method to assess the driver mental workload: the driving activity load index (DALI). *IET Intel Transp Syst* 2(4):315–322. <https://doi.org/10.1049/iet-its:20080023>
- Pollock E, Chandler P, Sweller J (2002) Assimilating complex information. *Learn Instr* 12(1):61–86. [https://doi.org/10.1016/S0959-4752\(01\)00016-0](https://doi.org/10.1016/S0959-4752(01)00016-0)
- Renkl A, Atkinson RK (2003) Structuring the transition from example study to problem solving in cognitive skill acquisition: a cognitive load perspective. *Educ Psychol* 38(1):15–22. https://doi.org/10.1207/S15326985EP3801_3
- Rose FD, Attree EA, Brooks BM, Parslow DM, Penn PR (2000) Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics* 43(4):494–511. <https://doi.org/10.1080/001401300184378>
- Shin D (2018) Empathy and embodied experience in virtual environment: to what extent can virtual reality stimulate empathy and embodied experience? *Comput Hum Behav* 78:64–73. <https://doi.org/10.1016/j.chb.2017.09.012>
- Somrak A, Humar I, Hossain MS, Alhamid MF, Hossain MA, Guna J (2019) Estimating VR sickness and user experience using different HMD technologies: an evaluation study. *Future Gener Comput Syst* 94:302–316. <https://doi.org/10.1016/j.future.2018.11.041>
- Stanney KM, Mollaghasemi M, Reeves L, Breau R, Graeber DA (2003) Usability engineering of virtual environments (VEs): identifying multiple criteria that drive effective VE system design. *Int J Hum Comput Stud* 58(4):447–481. [https://doi.org/10.1016/S1071-5819\(03\)00015-6](https://doi.org/10.1016/S1071-5819(03)00015-6)
- Steed A, Pan Y, Zisch F, Steptoe W (2016) The impact of a self-avatar on cognitive load in immersive virtual reality. *IEEE Virtual Real: VR* 2016:67–76. <https://doi.org/10.1109/VR.2016.7504689>
- Stevens JA, Kincaid JP (2015) The relationship between presence and performance in virtual simulation training. *Open J Model Simul* 03(02):41–48. <https://doi.org/10.4236/ojmsi.2015.32005>
- Sutherland LM, Middleton PF, Anthony A, Hamdorf J, Cregan P, Scott D, Maddern GJ (2006) Surgical simulation. *Ann Surg* 243(3):291–300. <https://doi.org/10.1097/01.sla.0000200839.93965.26>
- Sweller J (1999) Instructional design in technical areas. ACER Press, Melbourne, Australia
- R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/>
- Usoh M, Catena E, Arman S, Slater M (2000) Using presence questionnaires in reality. *Presence Teleoperators Virtual Environ* 9(5):497–503. <https://doi.org/10.1162/105474600566989>
- van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. Retrieved 10 Dec 2018 from <http://www.jstatsoft.org/v45/i03/>
- van Gog T, Ericsson KA, Rikers RMJP, Paas F (2005) Instructional design for advanced learners: establishing connections between the theoretical frameworks of cognitive load and deliberate practice. *Educ Tech Res Dev* 53(3):73–81. <https://doi.org/10.1007/BF02504799>
- Wickens CD (1992) Virtual reality and education. In: *Proceedings of 1992 IEEE international conference on systems, man, and cybernetics*, vol 1, pp 842–847. <https://doi.org/10.1109/ICSMC.1992.271688>
- Wilson MR, Poolton JM, Malhotra N, Ngo K, Bright E, Masters RSW (2011) Development and validation of a surgical workload measure: the surgery task load index (SURG-TLX). *World J Surg* 35(9):1961. <https://doi.org/10.1007/s00268-011-1141-4>
- Zijlstra FRH (1993) Efficiency in work behaviour: a design approach for modern tools. Retrieved 5 Dec 2018 from <http://resolver.tudelft.nl/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.